Data in Brief 6 (2016) 208–213

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data article

Curated eutherian third party data gene data sets

Marko Premzl

Laboratory of Genomics, Centre of Animal Reproduction, 55 Heinzel St., Zagreb, Croatia

ARTICLE INFO

Article history:
Received 5 September 2015
Received in revised form
18 November 2015
Accepted 24 November 2015
Available online 11 December 2015

Keywords:
Comparative Genomic Analysis
Gene Annotations
Molecular Evolution
Phylogenetic Analysis

ABSTRACT

The free available eutherian genomic sequence data sets advanced scientific field of genomics. Of note, future revisions of gene data sets were expected, due to incompleteness of public eutherian genomic sequence assemblies and potential genomic sequence errors. The eutherian comparative genomic analysis protocol was proposed as guidance in protection against potential genomic sequence errors in public eutherian genomic sequences. The protocol was applicable in updates of 7 major eutherian gene data sets, including 812 complete coding sequences deposited in European Nucleotide Archive as curated third party data gene data sets.

Specifications Table

Subject area	Biology
More specific subject area	Genomics
Type of data	Third party data
How data was acquired	*In computo*
Data format	FAS, TXT
Experimental factors	Eutherian comparative genomic analysis protocol
	Curated gene data sets

E-mail address: Marko.Premzl@alumni.anu.edu.au

http://dx.doi.org/10.1016/j.dib.2015.11.056

Experimental features	
Data source location	N/A
Data accessibility	The original gene data sets were deposited in European Nucleotide Archive under accession numbers: FR734011-FR734074 (http://www.ebi.ac.uk/ena/data/view/FR734011-FR734074), HF564658-HF564785 (http://www.ebi.ac.uk/ena/data/view/HF564658-HF564785), HF564786-HF564815 (http://www.ebi.ac.uk/ena/data/view/HF564786-HF564815), HG328835-HG329089 (http://www.ebi.ac.uk/ena/data/view/HG328835-HG329089), HG426065-HG426183 (http://www.ebi.ac.uk/ena/data/view/HG426065-HG426183), HG931734-HG931849 (http://www.ebi.ac.uk/ena/data/view/HG931734-HG931849) and LM644135-LM644234 (http://www.ebi.ac.uk/ena/data/view/LM644135-LM644234). Data analysis is with this article.

Value of the data

- Curated gene data sets applicable in gene annotations and genome analyses.
- Curated gene data sets applicable in phylogenetic analyses.
- Curated gene data sets applicable in protein structure and function analyses.

1. Data

Undoubtedly, the eutherian comparative genomics momentum was maintained by programmatic, considerable international efforts in production, assembly and analysis of public eutherian genomic sequence data sets (Fig. 1) [1–3]. For example, the initial sequencing and analysis of human genome revised human gene data sets [4,5]. Nevertheless, these analyses were subject to future updates and revisions due to incompleteness of public eutherian genomic sequence data sets and potential genomic sequence errors [1–6]. The eutherian comparative genomic analysis protocol was proposed as guidance in protection against potential genomic sequence errors in public eutherian genomic sequences [7–12]. The protocol was established as one framework of eutherian third party data gene data set descriptions (Fig. 2). The protocol included new genomics and protein molecular evolution tests applicable in updates and revisions of 7 major eutherian gene data sets, including interferon-γ-inducible GTPase genes, ribonuclease A genes, Mas-related G protein-coupled receptor genes, lysozyme genes, adenohypophysis cystine-knot genes, macrophage migration inhibitory factor and D-dopachrome tautomerase genes and, finally, growth hormone genes (Fig. 3). The protocol discriminated major gene clusters with and without evidence of differential gene expansions. For example, the eutherian major gene clusters with no evidence of differential gene expansions could be suitable in phylogenomic analyses.

2. Experimental design, materials and methods

The eutherian comparative genomic analysis protocol included gene annotations, phylogenetic analysis and protein molecular evolution analysis [7–12] (Fig. 2). The protocol used free available eutherian genomic sequence data sets deposited in public biological databases and software.

3. Gene annotations

The gene annotations included gene identifications in eutherian genomic sequences, analyses of gene features, tests of reliability of eutherian public genomic sequences and multiple pairwise

Superordinal clade	Order	Species
Euarchontoglires	Primates	*Homo sapiens* (Human)
		Pan troglodytes (Common chimpanzee)
		Gorilla gorilla (Western gorilla)
		Pongo abelii (Sumatran orangutan)
		Nomascus leucogenys (Northern white-cheeked gibbon)
		Macaca mulatta (Rhesus monkey)
		Papio hamadryas (Hamadryas baboon)
		Callithrix jacchus (Common marmoset)
		Tarsius syrichta (Philippine tarsier)
		Microcebus murinus (Gray mouse lemur)
		Otolemur garnettii (Northern greater galago)
	Scadentia	*Tupaia belangeri* (Northern treeshrew)
	Rodentia	*Mus musculus* (Mouse)
		Rattus norvegicus (Brown rat)
		Dipodomys ordii (Ord's kangaroo rat)
		Cavia porcellus (Domesticated guinea pig)
		Spermophilus tridecemlineatus (Thirteen-lined ground squirrel)
	Lagomorpha	*Oryctolagus cuniculus* (European rabbit)
		Ochotona princeps (American pika)
Laurasiatheria	Cetacea	*Tursiops truncatus* (Bottlenose dolphin)
	Artiodactyla	*Bos taurus* (Domestic cattle)
		Sus scrofa (Wild boar)
		Vicugna pacos (Vicugna)
	Perissodactyla	*Equus caballus* (Horse)
	Carnivora	*Canis lupus familiaris* (Domestic dog)
		Felis catus (Domestic cat)
	Chiroptera	*Myotis lucifugus* (Little brown myotis)
		Pteropus vampyrus (Large flying fox)
	Eulipotyphla	*Erinaceus europaeus* (West European hedgehog)
		Sorex araneus (Common shrew)
Xenarthra	Xenarthra	*Dasypus novemcinctus* (Nine-banded armadillo)
		Choloepus hoffmanni (Hoffmann's two-toed sloth)
Afrotheria	Tenrecidae	*Echinops telfairi* (Lesser hedgehog tenrec)
	Proboscidea	*Loxodonta africana* (African bush elephant)
	Hyracoidea	*Procavia capensis* (Rock hyrax)

Fig. 1. Public eutherian genomic sequence assemblies (http://www.ensembl.org).

genomic sequence alignments. The BioEdit program was used in nucleotide and protein sequence analyses (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The NCBI's BLAST programs were used in identifications of genes in eutherian genomic sequence assemblies downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/ and ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/). In addition, the Ensembl genome browser's BLAST or BLAT programs were used in gene identifications (http://www.ensembl.org). The analyses of gene features included direct evidence of eutherian gene annotations deposited in NCBI's nr, est_human, est_mouse and est_others databases (http://www.ncbi.nlm.nih.gov). The new tests of reliability of eutherian public genomic sequences tested potential coding sequences using genomic sequence redundancies. First, the tests analysed nucleotide sequence coverage of potential coding sequences using primary experimental sequence reads deposited in NCBI's Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) and BLAST programs. Second, the potential coding sequences were classified as complete coding sequences only if consensus trace sequence coverage was available for every nucleotide. Alternatively, the potential coding sequences were described as putative coding sequences. Only the complete coding sequences were deposited in European Nucleotide Archive as curated third party data gene data sets (http://www.ebi.ac.uk/ena/about/tpa-policy) and used in phylogenetic and protein molecular evolution analyses. In revised eutherian gene nomenclatures, the guidelines of human and mouse gene nomenclature were used (http://www.genenames.org/about/guidelines and http://www.informatics.jax.org/mgihome/nomen/gene.shtml). The maskings of transposable elements using RepeatMasker program were included as preparatory steps in multiple pairwise genomic sequence alignments (http://www.repeatmasker.org/). The RepeatMasker's default settings were used, except simple repeats and low complexity elements were not masked. The mVISTA program was used in genomic

sequence alignments, using AVID alignment algorithm and default settings (http://genome.lbl.gov/vista/index.shtml). Using ClustalW implemented in BioEdit, the common predicted promoter genomic sequence regions were aligned at nucleotide sequence level and then manually corrected. The pairwise nucleotide sequence identities of common predicted promoter genomic sequence regions calculated using BioEdit were used in statistical analyses (Microsoft Office Excel).

4. Phylogenetic analysis

The phylogenetic analyses included protein and nucleotide sequence alignments, calculations of phylogenetic trees and calculations of pairwise nucleotide sequence identity patterns. First, the translated complete coding sequences were aligned at amino acid level using ClustalW implemented

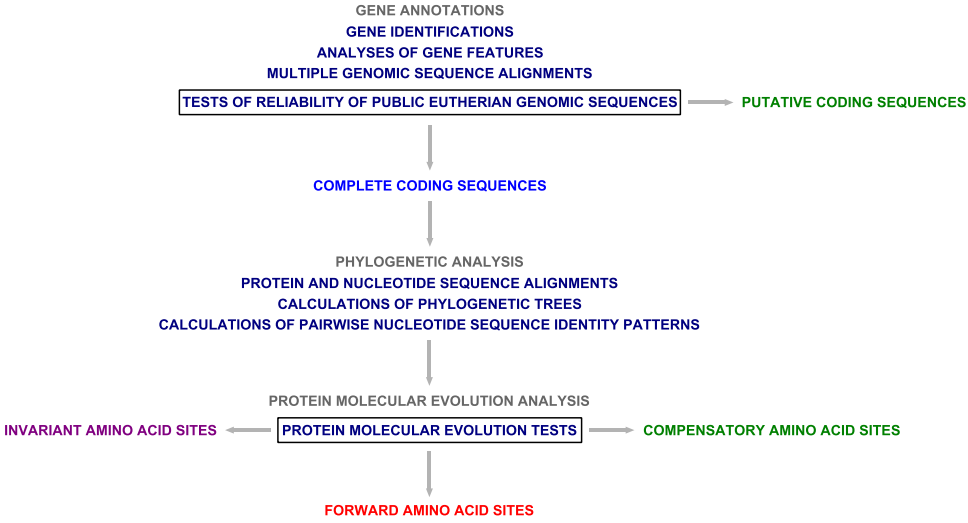

Fig. 2. Eutherian comparative genomic analysis protocol scheme.

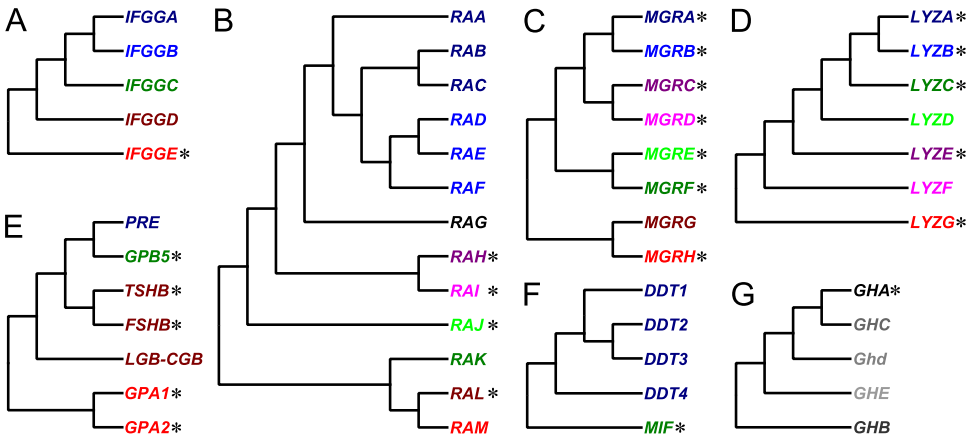

Fig. 3. Revised gene classifications of eutherian interferon-γ-inducible GTPase genes (A), ribonuclease A genes (B), Mas-related G protein-coupled receptor genes (C), lysozyme genes (D), adenohypophysis cystine-knot genes (E) and growth hormone genes (G) and human D-dopachrome tautomerase and macrophage migration inhibitory factor genes (F). The major gene clusters with no evidence of differential gene expansions were indicated by *s.

in BioEdit. The protein sequence alignments were manually corrected, as well as nucleotide sequence alignments. The MEGA program was used in phylogenetic tree calculations (http://www.megasoftware.net), using neighbour-joining method (default settings, except gaps/missing data treatment=pairwise deletion), minimum evolution method (default settings, except gaps/missing data treatment=pairwise deletion) and maximum parsimony method (default settings, except gaps/missing data treatment=use all sites). The pairwise nucleotide sequence identities of complete coding sequences were calculated using BioEdit and used in statistical analysis (Microsoft Office Excel).

5. Protein molecular evolution analysis

The protocol included new protein molecular evolution tests integrating patterns of nucleotide sequence similarities with protein tertiary structures. The MEGA program was used in calculations of codon usage statistics. Specifically, the ratios between observed and expected amino acid codon counts determined relative synonymous codon usage statistics (R) that indicated amino acid codons with $R \leq 0.7$ as not preferable amino acid codons. In reference protein amino acid sequences, there were invariant amino acid sites (invariant alignment positions), forward amino acid sites (variant alignment positions that did not include not preferable amino acid codons) and compensatory amino acid sites (variant alignment positions that included not preferable amino acid codons). The presence of preferable amino acid codons, as well as absence of not preferable amino acid codons indicated that forward amino acid sites could have major influence on protein tertiary structures and functions. The DeepView/Swiss-PdbViever was used in analyses of protein tertiary structures (http://spdbv.vital-it.ch/).

Acknowledgements

The author would like to thank publisher on discounted article publication fees.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2015.11.056.

References

[1] R.W. Blakesley, N.F. Hansen, J.C. Mullikin, P.J. Thomas, J.C. McDowell, B. Maskeri, A.C. Young, B. Benjamin, S.Y. Brooks, B. I. Coleman, J. Gupta, S.L. Ho, E.M. Karlins, Q.L. Maduro, S. Stantripop, C. Tsurgeon, J.L. Vogt, M.A. Walker, C.A. Masiello, X. Guan, NISC Comparative Sequencing Program, G.G. Bouffard, E.D. Green, An intermediate grade of finished genomic sequence suitable for comparative analyses, Genome Res. 14 (2004) 2235–2244.

[2] E.H. Margulies, J.P. Vinson, NISC Comparative Sequencing Program, W. Miller, D.B. Jaffe, K. Lindblad-Toh, J.L. Chang, E. D. Green, E.S. Lander, J.C. Mullikin, M. Clamp, An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing, Proc. Natl. Acad. Sci. USA 102 (2005) 4795–4800.

[3] K. Lindblad-Toh, M. Garber, O. Zuk, M.F. Lin, B.J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L.D. Ward, C.B. Lowe, A.K. Holloway, M. Clamp, S. Gnerre, J. Alföldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M.J. Hubisz, D.B. Jaffe, I. Jungreis, W.J. Kent, D. Kostka, M. Lara, A.L. Martins, T. Massingham, I. Moltke, B.J. Raney, M.D. Rasmussen, J. Robinson, A. Stark, A.J. Vilella, J. Wen, X. Xie, M.C. Zody, Broad Institute Sequencing Platform and Whole Genome Assembly Team, J. Baldwin, T. Bloom, C.W. Chin, D. Heiman, R. Nicol, C. Nusbaum, S. Young, J. Wilkinson, K.C. Worley, C.L. Kovar, D.M. Muzny, R.A. Gibbs, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, A. Cree, H.H. Dihn, G. Fowler, S. Jhangiani, V. Joshi, S. Lee, L.R. Lewis, L.V. Nazareth, G. Okwuonu, J. Santibanez, W.C. Warren, E.R. Mardis, G.M. Weinstock, R.K. Wilson, Genome Institute at Washington University, K. Delehaunty, D. Dooling, C. Fronik, L. Fulton, B. Fulton, T. Graves, P. Minx, E. Sodergren, E. Birney, E.H. Margulies, J. Herrero, E.D. Green, D. Haussler, A. Siepel, N. Goldman, K.S. Pollard, J.S. Pedersen, E.S. Lander, M. Kellis, A high-resolution map of human evolutionary constraint using 29 mammals, Nature 478 (2011) 476–482.

[4] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[5] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T.J. Hubbard, GENCODE: the reference human genome annotation for the ENCODE project, Genome. Res. 22 (2012) 1760–1774.
[6] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[7] M. Premzl, Comparative genomic analysis of eutherian interferon-γ-inducible GTPases, Funct. Integr. Genom. 12 (2012) 599–607.
[8] M. Premzl, Comparative genomic analysis of eutherian ribonuclease A genes, Mol. Genet. Genom. 289 (2014) 161–167.
[9] M. Premzl, Comparative genomic analysis of eutherian Mas-related G protein-coupled receptor genes, Gene 540 (2014) 16–19.
[10] M. Premzl, Third party annotation gene data set of eutherian lysozyme genes, Genom. Data 2 (2014) 258–260.
[11] M. Premzl, Initial description of primate-specific cystine-knot Prometheus genes and differential gene expansions of D-dopachrome tautomerase genes, Meta Gene 4 (2015) 118–128.
[12] M. Premzl, Third party data gene data set of eutherian growth hormone genes, Genom. Data 6 (2015) 166–169.